Notes on Matrix Derivative and Trace

Wu Xiaojian

July 27,2019

1 Notation

Let

$$f : \mathbb{R}^{m\times n} \mapsto \mathbb{R} \tag{1}$$

be a scalar function, taking matrix of $m \times n$ shape as input, and throwing out a real number as output. By convention, we use bold capitalized letters such as $\boldsymbol{X}, \boldsymbol{Y}$ to denote matrix, as follows

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}$$

We define the derivative of f with respect to $\boldsymbol{X}$ as

$$\frac{\partial f}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial f}{\partial x_{1,1}} & \frac{\partial f}{\partial x_{1,2}} & \cdots & \frac{\partial f}{\partial x_{1,n}} \\ \frac{\partial f}{\partial x_{2,1}} & \frac{\partial f}{\partial x_{2,2}} & \cdots & \frac{\partial f}{\partial x_{2,n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m,1}} & \frac{\partial f}{\partial x_{m,2}} & \cdots & \frac{\partial f}{\partial x_{m,n}} \end{bmatrix} \tag{2}$$

that is, the derivative applies to each elements of $\boldsymbol{X}$ and is still a matrix with the same shape as $\boldsymbol{X}$, much like to the broadcast operation upon a matrix which applies the specified operation to all the elements and remains the shape. Similarly, we define the differential of $\boldsymbol{X}$ as

$$\mathrm{d}\boldsymbol{X} = \begin{bmatrix} \mathrm{d}x_{1,1} & \mathrm{d}x_{1,2} & \cdots & \mathrm{d}x_{1,n} \\ \mathrm{d}x_{2,1} & \mathrm{d}x_{2,2} & \cdots & \mathrm{d}x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{d}x_{m,1} & \mathrm{d}x_{m,2} & \cdots & \mathrm{d}x_{m,n} \end{bmatrix} \tag{3}$$

Finally, let's introduce *trace*. For a square $\boldsymbol{X}^{n\times n}$, we use $\mathbf{tr}(\boldsymbol{X})$ to represent the trace of $\boldsymbol{X}$, the sum of those among in the main diagonal of $\boldsymbol{X}$

$$\mathbf{tr}(\boldsymbol{X}) = \sum_{i=1}^{n} x_{i,i} \tag{4}$$

which shows[1] $\mathbf{tr}(\boldsymbol{X}) \in \mathbb{R}$.

2 Inner Product of Matrix

For any $\boldsymbol{v}, \boldsymbol{k} \in \mathbb{R}^n$,their inner product can be defined as

$$\langle \boldsymbol{v}, \boldsymbol{k} \rangle = \boldsymbol{v}^{\mathrm{T}} \boldsymbol{k} = \sum_{i=1}^{n} v_i k_i$$

It's very natural to borrow this concept and define the *inner product of matrix*. Suppose $\boldsymbol{A}, \boldsymbol{B}$ are matrix of the shame shape $m \times n$,their inner product is

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{m \times n} = \sum_{\substack{1 \leqslant j \leqslant m \\ 1 \leqslant i \leqslant n}} a_{i,j} b_{i,j}$$

that is, the sum of product entrywise of two matrix. Now that we have defined trace, we can use trace to define inner product of matrix, in a elegant way.

Theorem 2.1 *Suppose $\boldsymbol{A}, \boldsymbol{B}$ are matrix of the shame shape $m \times n$, then*

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathbf{tr}(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{B}) \tag{5} \quad \square$$

Proof *Suppose $\boldsymbol{A} = [\alpha_1, \alpha_2, \ldots, \alpha_n]$, $\boldsymbol{B} = [\beta_1, \beta_2, \ldots, \beta_n]$ where α_i, β_i are column vectors, then $\boldsymbol{A}^{\mathrm{T}} = [\alpha_1^{\mathrm{T}}, \alpha_2^{\mathrm{T}}, \cdots, \alpha_n^{\mathrm{T}}]^{\mathrm{T}}$, and*

$$A^{\mathrm{T}} B = \begin{bmatrix} \alpha_1^{\mathrm{T}} \\ \alpha_2^{\mathrm{T}} \\ \vdots \\ \alpha_n^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \beta_1, \beta_2, \cdots, \beta_n \end{bmatrix} = \begin{bmatrix} \alpha_1^{\mathrm{T}} \beta_1 & & & \\ & \alpha_2^{\mathrm{T}} \beta_2 & & \\ & & \ddots & \\ & & & \alpha_n^{\mathrm{T}} \beta_n \end{bmatrix}$$

the other entries of which are not important and can be omitted. Then

$$\mathbf{tr}(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{B}) = \sum_{j=1}^{n} \alpha_j^{\mathrm{T}} \beta_j = \sum_{j=1}^{n} \sum_{i=1}^{m} a_{ij} b_{ij} = \langle \boldsymbol{A}, \boldsymbol{B} \rangle$$

3 Derivative of Multi-variable Function

Recall that in multi-variable caluculus, the *complete differential* of f with respect to variables $x_1, x_2, \cdots, x_n$ is

$$\mathrm{d}f = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \mathrm{d}x_i$$

[1] We assume all matrix we discussing here are real matrix.

Using nabla operator $\nabla = [\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \cdots, \frac{\partial}{\partial x_n}]^{\mathrm{T}}$, we rewrite the complete differential above as more gracefully as

$$\mathrm{d}f = \nabla f \cdot \mathrm{d}\boldsymbol{x} = (\nabla f)^{\mathrm{T}} \mathrm{d}\boldsymbol{x} \tag{6}$$

where

$$\nabla f = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}]^{\mathrm{T}} \quad \text{and} \quad \mathrm{d}\boldsymbol{x} = [\mathrm{d}x_1, \mathrm{d}x_2, \cdots, \mathrm{d}x_n]^{\mathrm{T}}$$

Now we can generalize this concept to matrix, say, the complete differential of f with respect to $\boldsymbol{X}_{m\times n}$, and so we have

$$\mathrm{d}f\big|_{\boldsymbol{X}} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial f}{\partial x_{i,j}} \mathrm{d}x_{i,j} \tag{7}$$

in most cases, if context agrees, we can denote it shortly as $\mathrm{d}f$.

Example 3.1 *Suppose*

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix}$$

we have

$$\frac{\partial f}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \frac{\partial f}{\partial x_{13}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \frac{\partial f}{\partial x_{23}} \end{bmatrix} \quad \textit{and} \quad \mathrm{d}\boldsymbol{X} = \begin{bmatrix} \mathrm{d}x_{11} & \mathrm{d}x_{12} & \mathrm{d}x_{13} \\ \mathrm{d}x_{21} & \mathrm{d}x_{22} & \mathrm{d}x_{23} \end{bmatrix}$$

Let's transpose the derivative and multiply them, and we can get

$$\left(\frac{\partial f}{\partial \boldsymbol{X}}\right)^{\mathrm{T}} \mathrm{d}\boldsymbol{X} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}}\mathrm{d}x_{11} + \frac{\partial f}{\partial x_{21}}\mathrm{d}x_{21} & & \\ & \frac{\partial f}{\partial x_{12}}\mathrm{d}x_{12} + \frac{\partial f}{\partial x_{22}}\mathrm{d}x_{22} & \\ & & \frac{\partial f}{\partial x_{13}}\mathrm{d}x_{13} + \frac{\partial f}{\partial x_{23}}\mathrm{d}x_{23} \end{bmatrix}$$

It's very obvious that

$$\mathbf{tr}\left(\left(\frac{\partial f}{\partial \boldsymbol{X}}\right)^{\mathrm{T}} \mathrm{d}\boldsymbol{X}\right) = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{\partial f}{\partial x_{ij}} \mathrm{d}x_{ij} = \mathrm{d}f \tag{8}$$

In Example(3.1) we find it seems that the trace of the multiplication between the transpose of derivative of f with respect to $\boldsymbol{X}$ and the differential of $\boldsymbol{X}$ equals to the differential of f. In fact, we have the theorem, as follows

Theorem 3.1 *Suppose* $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ *and* $f : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$*, then*

$$\mathrm{d}f = \langle \frac{\partial f}{\partial \boldsymbol{X}}, \mathrm{d}\boldsymbol{X} \rangle = \mathbf{tr}\left(\left(\frac{\partial f}{\partial \boldsymbol{X}}\right)^{\mathrm{T}} \mathrm{d}\boldsymbol{X}\right) \tag{9} \quad \square$$

Proof *By* THEOREM*(2.1), we have*

$$\langle \frac{\partial f}{\partial \boldsymbol{X}}, \mathrm{d}\boldsymbol{X} \rangle = \mathbf{tr}\left(\left(\frac{\partial f}{\partial \boldsymbol{X}} \right)^{\mathrm{T}} \mathrm{d}\boldsymbol{X} \right)$$

On the other hand, since the inner product of two matrix is the sum of each product entrywise of the two, that is,

$$\langle \frac{\partial f}{\partial \boldsymbol{X}}, \mathrm{d}\boldsymbol{X} \rangle = \sum_{i,j} \frac{\partial f}{\partial x_{i,j}} \mathrm{d}x_{i,j}$$

we have $\mathrm{d}f = \langle \frac{\partial f}{\partial \boldsymbol{X}}, \mathrm{d}\boldsymbol{X} \rangle$*, and thus the theorem holds.* ■

4 Matrix Differential

Analogous to laws of derivative on real-valued function, we can define the derivative of matrix and derive some operation laws from it.

Theorem 4.1 (Basic Laws of Arithmetic Operations)

$$\mathrm{d}(\boldsymbol{X} \pm \boldsymbol{Y}) = \mathrm{d}\boldsymbol{X} \pm \mathrm{d}\boldsymbol{Y} \tag{10}$$

$$\mathrm{d}(\boldsymbol{X}\boldsymbol{Y}) = (\mathrm{d}\boldsymbol{X})\boldsymbol{Y} + \boldsymbol{X}(\mathrm{d}\boldsymbol{Y}) \tag{11}$$

$$\mathrm{d}(\boldsymbol{X} \odot \boldsymbol{Y}) = \mathrm{d}(\boldsymbol{X}) \odot \boldsymbol{Y} + \boldsymbol{X} \odot \mathrm{d}(\boldsymbol{Y}) \tag{12}$$

where $\odot$ *is* ***Hadamard Product****(also known as the* ***Schur product*** *or the* ***Entrywise Product****) is a binary operation that takes two matrices of the same dimensions, and produces another matrix where each element ij is the product of elements ij of the original two matrices. It should not be confused with the more common matrix product.* □

Proof

$$\begin{aligned}
\mathrm{d}(\boldsymbol{X} \pm \boldsymbol{Y}) &= \begin{bmatrix} \mathrm{d}(x_{11} \pm y_{11}) & \cdots & \mathrm{d}(x_{1n} \pm y_{1n}) \\ \vdots & \ddots & \vdots \\ \mathrm{d}(x_{m1} \pm y_{m1}) & \cdots & \mathrm{d}(x_{mn} \pm y_{mn}) \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{d}x_{11} \pm \mathrm{d}y_{11} & \cdots & \mathrm{d}x_{1n} \pm \mathrm{d}y_{1n} \\ \vdots & \ddots & \vdots \\ \mathrm{d}x_{m1} \pm \mathrm{d}y_{m1} & \cdots & \mathrm{d}x_{mn} \pm \mathrm{d}y_{mn} \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{d}x_{11} & \cdots & \mathrm{d}x_{1n} \\ \vdots & \ddots & \vdots \\ \mathrm{d}x_{m1} & \cdots & \mathrm{d}x_{mn} \end{bmatrix} \pm \begin{bmatrix} \mathrm{d}y_{11} & \cdots & \mathrm{d}y_{1n} \\ \vdots & \ddots & \vdots \\ \mathrm{d}y_{m1} & \cdots & \mathrm{d}y_{mn} \end{bmatrix} \\
&= \mathrm{d}\boldsymbol{X} \pm \mathrm{d}\boldsymbol{Y}
\end{aligned}$$

$$\mathrm{d}(\boldsymbol{XY}) = \mathrm{d}\left(\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{n2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} y_{11} & \cdots & y_{1t} \\ y_{21} & \cdots & y_{2t} \\ \vdots & \cdots & \vdots \\ y_{n1} & \cdots & y_{nt} \end{bmatrix}\right) = \mathrm{d}\begin{bmatrix} z_{11} & \cdots & z_{1t} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nt} \end{bmatrix}$$

where $z_{ij} = \sum_{k=1}^{n} x_{ik} y_{kj}$*, so*

$$\mathrm{d}z_{ij} = \mathrm{d}\sum_{k=1}^{n} x_{ik} y_{kj} = \sum_{k=1}^{n} \mathrm{d}(x_{ik}) y_{kj} + \sum_{k=1}^{n} x_{ik} \mathrm{d}(y_{kj})$$

For the last two items, we can convert each of them back to the form of matrix multiplication and so we will get

$$\mathrm{d}\boldsymbol{Z} = \mathrm{d}(\boldsymbol{X})\boldsymbol{Y} + \boldsymbol{X}\mathrm{d}(\boldsymbol{Y})$$

The proof of the hadamard product is similar. ■

Theorem 4.2

$$\mathrm{d}(\boldsymbol{X}^{\mathrm{T}}) = (\mathrm{d}\boldsymbol{X})^{\mathrm{T}} \tag{13}$$

$$\mathrm{d}(\mathbf{tr}(\boldsymbol{X})) = \mathbf{tr}(\mathrm{d}\boldsymbol{X}) \tag{14}$$

Theorem 4.3 (Differential of Constant Matrix) *Assume* $\boldsymbol{C}$ *is a constant matrix, then*

$$\mathrm{d}(\boldsymbol{C}) = O \tag{15}$$ □

Theorem 4.4 (Differential of Inverse of Matrix) *Suppose* $\boldsymbol{X}$ *is invertable, and its inverse is* $\boldsymbol{X}^{-1}$*, then*

$$\mathrm{d}(\boldsymbol{X}^{-1}) = -\boldsymbol{X}^{-1}\mathrm{d}(\boldsymbol{X})\boldsymbol{X}^{-1} \tag{16}$$ □

Proof *Let* $\boldsymbol{I}$ *be identity.Since* $\boldsymbol{X}$ *is invertable, therefore* $\mathrm{d}\boldsymbol{I} = \mathrm{d}(\boldsymbol{X}\boldsymbol{X}^{-1})$*. Notice* $\boldsymbol{I}$ *is a constant matrix, thus* $\mathrm{d}\boldsymbol{I} = O$ *and*

$$\mathrm{d}(\boldsymbol{X}\boldsymbol{X}^{-1}) = \mathrm{d}(\boldsymbol{X})\boldsymbol{X}^{-1} + \boldsymbol{X}\mathrm{d}(\boldsymbol{X}^{-1}) = O$$

This concludes that $\mathrm{d}(\boldsymbol{X}^{-1}) = -\boldsymbol{X}^{-1}\mathrm{d}(\boldsymbol{X})\boldsymbol{X}^{-1}$*.* ■

Theorem 4.5 (Chain Rule) *Let* σ *is a scalar function, and we define the broadcast operation as*

$$\sigma(\boldsymbol{X}) = \begin{bmatrix} \sigma(x_{11}) & \cdots & \sigma(x_{1n}) \\ \vdots & \ddots & \vdots \\ \sigma(x_{m1}) & \cdots & \sigma(x_{mn}) \end{bmatrix}$$

then

$$\mathrm{d}\left(\sigma(\boldsymbol{X})\right) = \sigma'(\boldsymbol{X}) \odot \mathrm{d}\boldsymbol{X} \tag{17}$$ □

Proof

$$
\begin{aligned}
\mathrm{d}\left(\sigma(\boldsymbol{X})\right) &= \begin{bmatrix} \mathrm{d}\sigma(x_{11}) & \cdots & \mathrm{d}\sigma(x_{1n}) \\ \vdots & \ddots & \vdots \\ \mathrm{d}\sigma(x_{m1}) & \cdots & \mathrm{d}\sigma(x_{mn}) \end{bmatrix} \\
&= \begin{bmatrix} \sigma'(x_{11})\mathrm{d}x_{11} & \cdots & \sigma'(x_{1n})\mathrm{d}x_{1n} \\ \vdots & \ddots & \vdots \\ \sigma'(x_{m1})\mathrm{d}x_{m1} & \cdots & \sigma'(x_{mn})\mathrm{d}x_{mn} \end{bmatrix} \\
&= \begin{bmatrix} \sigma'(x_{11}) & \cdots & \sigma'(x_{1n}) \\ \vdots & \ddots & \vdots \\ \sigma'(x_{m1}) & \cdots & \sigma'(x_{mn}) \end{bmatrix} \odot \begin{bmatrix} \mathrm{d}x_{11} & \cdots & \mathrm{d}x_{1n} \\ \vdots & \ddots & \vdots \\ \mathrm{d}x_{m1} & \cdots & \mathrm{d}x_{mn} \end{bmatrix} \\
&= \sigma'(\boldsymbol{X}) \odot \mathrm{d}\boldsymbol{X}
\end{aligned}
$$

5 Trace

Theorem 5.1 *Suppose a, b, c is scalar value, $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m\times n}$, $\boldsymbol{Z} \in \mathbb{R}^{n\times m}$. Here are some properties about trace*

$$\mathbf{tr}(c) = c \tag{18}$$

$$\mathbf{tr}(\boldsymbol{X}^{\mathrm{T}}) = \mathbf{tr}(\boldsymbol{X}) \tag{19}$$

$$\mathbf{tr}(a\boldsymbol{X} \pm b\boldsymbol{Y}) = a\mathbf{tr}(\boldsymbol{X}) \pm b\mathbf{tr}(\boldsymbol{Y}) \tag{20}$$

$$\mathbf{tr}(\boldsymbol{X}\boldsymbol{Z}) = \mathbf{tr}(\boldsymbol{Z}\boldsymbol{X}) \tag{21}$$

Proof *we just proof the last one. Since the (i,j)-entry of $\boldsymbol{X}\boldsymbol{Z}$ is $\sum_{k=1}^{n} x_{ik}z_{kj}$, and that of $\boldsymbol{Z}\boldsymbol{X}$ is $\sum_{p=1}^{m} z_{ip}x_{pj}$, thus*

$$\mathbf{tr}(\boldsymbol{X}\boldsymbol{Z}) = \sum_{p=1}^{m}\sum_{k=1}^{n} x_{pk}z_{kp} = \sum_{k=1}^{n}\sum_{p=1}^{m} z_{kp}x_{pk} = \mathbf{tr}(\boldsymbol{Z}\boldsymbol{X})$$

We can also infer from this property that if $\boldsymbol{X}_1\boldsymbol{X}_2\cdots\boldsymbol{X}_n$ and $\boldsymbol{X}_n\boldsymbol{X}_1\cdots\boldsymbol{X}_{n-1}$ are both defined, then

$$\mathbf{tr}(\boldsymbol{X}_1\boldsymbol{X}_2\cdots\boldsymbol{X}_n) = \mathbf{tr}(\boldsymbol{X}_n\boldsymbol{X}_1\cdots\boldsymbol{X}_{n-1}) \tag{22}$$

The proof is very easy that using $\mathbf{tr}(\boldsymbol{X}\boldsymbol{Z}) = \mathbf{tr}(\boldsymbol{Z}\boldsymbol{X})$, let $\boldsymbol{X} = \boldsymbol{X}_1\boldsymbol{X}_2\cdots\boldsymbol{X}_{n-1}$ and $\boldsymbol{Z} = \boldsymbol{X}_n$ and so we make it. ■